Web Data Extractors Sites and Resources

By

Marcus P. Zillman, M.S., A.M.H.A.
Executive Director – Virtual Private Library
zillman@virtualprivatelibrary.com

This July 2009 column **Web Data Extractors Sites and Resources** is a comprehensive list of resources and sites that give you the latest and most important information concerning web data gathering and extractors that are available over the Internet including related and associated resources and sites. The below list is taken from my Web Data Extractors White Paper and is constantly updated with Subject Tracer™ bots at the following URL:

http://www.WebDataExtractors.com/

These sites and resources allow you to gather data from the world wide web using your computer, netbook, cell phone, or mobile device! Gathering data for research on the Internet is a must in today's highly competitive business environment and these resources will help you build a comprehensive database of niched research information taken directly from the web.

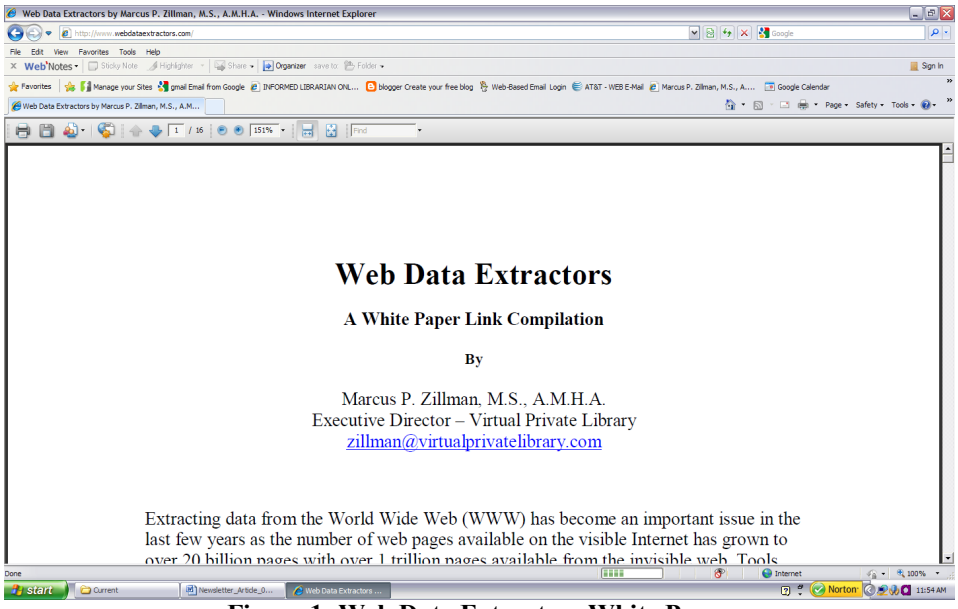

Figure 1: Web Data Extractors White Paper

Web Data Extractors

A White Paper Link Compilation

By

Marcus P. Zillman, M.S., A.M.H.A.
Executive Director – Virtual Private Library
zillman@virtualprivatelibrary.com

Extracting data from the World Wide Web (WWW) has become an important issue in the last few years as the number of web pages available on the visible Internet has grown to over 20 billion pages with over 1 trillion pages available from the invisible web. Tools and protocols to extract all this information have now come in demand as researchers as well as web browsers and surfers want to discovery new knowledge at an ever increasing rate! As robots (bots) and intelligent agents are at the heart of many extraction tools I decided to create a compilation of the latest sources and sites that extract information from the web. There are a number of eMail extraction tools still available through the Internet and I have decided not to list these as they aid to the on-going and increasing problem of SPAM except for a readily available Google™ Directory listing:

Web Data Extractors:

Anomic HTTP Proxy
http://freshmeat.net/releases/173068/

Anthracite
http://freshmeat.net/projects/anthracite/

Automated RSS Scraper Scripts
http://www.djeaux.com/rss/

Automated Information Solutions
http://www.automated-info-solutions.com/

Automatic Information Extraction From Semi-Structured Web Pages By Pattern Discovery
http://portal.acm.org/citation.cfm?id=640423&dl=ACM&coll=portal

Beautiful Soup
http://freshmeat.net/projects/beautifulsoup/

BLIASoft Knowledge Discovery
http://www.bliasoft.com/Eindex.html

Bot Research
http://www.BotResearch.info/

BYU Data Extraction Research Group
http://www.deg.byu.edu/

Captiva Software: Digital Information Capture Software
http://www.captivasoftware.com/index.asp

ChartSearch Data Search Technology
http://www.ChartSearch.net/

Client-Side Deep Web Data Extraction
http://www.tic.udc.es/~mad/publications/ceceast2004.pdf

Connotate - Intelligent Agent Technology and Business Intelligence Tools
http://www.connotate.com/intelligent_software_agents.aspx

Dapper - Extract and Use Website Information for Mashups
http://www.dappit.com/index.php

Data Extractors and Pass-Through Systems – A Selected List
http://www.chass.utoronto.ca/datalib/misc/dli/extracts.htm

Data Extractors – Special Report
http://snipurl.com/8flw

Data Mining Resources
http://www.DataMiningResources.info/

Deep Web Research
http://www.DeepWebResearch.info/

Effective Web Data Extraction with Standard XML Technologies
http://www10.org/cdrom/papers/102/

ExtractData Technologies - SearchExtract Software
http://www.extradata.com/

Extracting Knowledge
http://www.intelligentkm.com/feature/010507/feat1.jhtml?_requestid=185742

FerretSoft
http://www.FerretSoft.com/

Ficstar Software - Web Data Extraction
http://www.ficstar.com/index.html

Google™ Directory – Extractors
http://directory.google.com/Top/Computers/Software/Shareware/Windows/Internet/Email/Extractors/

Happy Harvester – Advanced Web Data Extractor
http://www.happyharvester.com/

Imagination Engines
http://www.Imagination-Engines.com/

Information Retrieval (IR) and Information Extraction (IE) on the Web
http://www.webir.org/

Information Retrieval Resources
http://www-csli.stanford.edu/~hinrich/information-retrieval.html

InfoSquire – Web Data Extraction
http://www.InfoSquire.com/

Introduction to Information Retrieval
http://informationretrieval.org/

iOpus® Internet Macros™
http://www.iopus.com/iim.htm

iWebMiner – Web Data Mining and Content Scraping
http://www.iwebminer.com/

Kapow RoboSuite Platform - Solutions for Data Collection
http://www.kapowtech.com/

Knowledge Discovery Resources
http://www.KnowledgeDiscovery.info/

Knowlesys® - Web Data Extraction, Web Grabber and Screen Scraper
http://www.knowlesys.com/index.htm

LingPipe – Information Extraction and Data Mining Tools
http://alias-i.com/lingpipe/

Lingway
http://www.lingway.fr/en/

Mastering Data Extraction
http://www.dbmsmag.com/9606d05.html

Metadata Extraction Tool
http://meta-extractor.sourceforge.net/

mnot: xpath2rss - HTML->RSS scraper
http://www.mnot.net/xpath2rss/

Mozenda – Comprehensive Web Data Gathering
http://www.mozenda.com/

NewsClipper.com - Snip and Ship Dynamic News Content to Your Web Pages
http://www.newsclipper.com/

Pervasive Data Management and Integration Products
http://www.pervasive.com/

Professional Web Data Extraction and Web Data Mining Solutions and Services
http://www.web-extraction.com/

QL2 Software - Unstructured Data Management and Web Mining Software
http://www.ql2.com/

REBOL Technologies
http://www.rebol.com/

ScrapeForge
http://freshmeat.net/projects/scrapeforge/

ScrapeGoat
http://www.ScrapeGoat.com/

Scraper
http://freshmeat.net/projects/scraper/

Screen-Scraper
http://freshmeat.net/projects/screenscraper/

Screen-Scraper – Extracts Information From Web Sites
http://www.Screen-Scraper.com/

Screenscraping the Senate by Paul Ford
http://www.xml.com/pub/a/2004/09/01/hack-congress.html

Screen Snarfs - Darkspell: Perl Code for Screen Scrapers
http://www.darkspell.com/gadgets/snarfs/

Search and Replace with TextPipe Pattern Matching
http://www.crystalsoftware.com.au/textpipe.html

Social Science Data Extractors
http://www.ssc.wisc.edu/cde/datalib/extract.htm

Text Mining and Web-Based Information Retrieval Reference
http://filebox.vt.edu/users/wfan/text_mining.html

Unit Miner - Web Data Extraction Software
http://www.qualityunit.com/unitminer/web-extraction-tool.html

URL Link Extractor
http://www.fuddyduddy.connectfree.co.uk/urlgen.htm

Visual Web Spider
http://www.newprosoft.com/

Visual Web Task
http://www.lencom.com/VisualWTSite.html

W3C Publishes Data Extraction Language (DEL) as W3C Note
http://xml.coverpages.org/ni2001-11-06-a.html

Web Data Extraction Software
http://www.tethyssolutions.com/web-data-extraction.htm

Web Data Extractor
http://www.webextractor.com/

Web Data Extractor
http://www.rafasoft.com/

Web Data Mining
http://www.blossom.com/web_mining.html

Web Grabber
http://www.ficstar.com/web_grabber.html

Web-Harvest – Open Source Web Data Extraction Tool
http://web-harvest.sourceforge.net/index.php

Web Mining and Unstructured Data Management Solutions – QL2 Software
http://www.ql2.com/

WebQL
http://www.ig.com.au

WebScraper Plus +
http://www.velocityscape.com/

Website Extractor
http://www.hot-shareware.com/internet-tools/website-extractor/

Website Extractor – Offline Browser
http://www.internet-soft.com/extractor.htm

Website Scraping
http://www.websitescraping.com/

Web Spider, Link Extraction, And Other Extractor Products
http://www.pjltechnology.com/

Words,Extended - Internet Text Information Rretrieval, Extraction and Display Bot
http://home.earthlink.net/~glenn_scheper/

XRay Web Scraping Tool
http://freshmeat.net/projects/xrayguibasedwebscrapingtool/

Subject Tracer™ Information Blogs

Subject Tracer™ Information Blogs created and developed by the Virtual Private Library™ combine the best of the latest tools on the Internet. Using bots, blogs and news aggregators the Subject Tracer™ Information blogs generate RSS feeds with the latest resources to create a current information resource flow through niched subject tracers. I am proud to be the creator of the Internet's first Subject Tracer™ Information Blogs:

Virtual Private Library™
http://www.VirtualPrivateLibrary.com/

Agriculture Resources
http://www.AgricultureResources.info/

Artificial Intelligence Resources
http://www.AIResources.info/

Astronomy Resources
http://www.AstronomyResources.info/

Auction Resources
http://www.AuctionResources.info/

Biological Informatics
http://www.BiologicalInformatics.info/

Biotechnology Resources
http://www.BiotechnologyResources.info/

Bot Research
http://www.BotResearch.info/

Business Intelligence Resources
http://www.BIResources.info/

ChatterBots
http://www.ChatterBots.info/

Data Mining Resources
http://www.DataMiningResources.info/

Deep Web Research
http://www.DeepWebResearch.info/

Directory Resources
http://www.DirectoryResources.info/

eCommerce Resources
http://eCommerceResources.info/

Elder Resources
http://www.ElderResources.info/

Employment Resources
http://www.EmploymentResources.info/

Entrepreneurial Resources
http://www.EntrepreneurialResources.info/

Financial Sources
http://www.FinancialSources.info/

Finding People
http://www.FindingPeople.info/

Games Resources
http://www.GamesResources.info/

Genealogy Resources
http://www.GenealogyResources.info/

Grant Resources
http://www.GrantResources.info/

Green Files
http://www.GreenFiles.info/

Grid, Distributed and Cloud Computing Resources
http://www.GridResources.info/

Healthcare Resources
http://www.HealthcareResources.info/

Information Futures Markets
http://www.InformationFutureMarkets.com/

Information Quality Resources
http://www.InformationQualityResources.info/

International Trade Resources
http://www.InternationalTradeResources.info/

Internet Alerts
http://www.InternetAlerts.info/

Internet Demographics
http://www.InternetDemographics.info/

Internet Experts
http://www.InternetExperts.info/

Internet Hoaxes
http://www.InternetHoaxes.info/

Journalism Resources
http://www.JournalismResources.info/

Knowledge Discovery
http://www.KnowledgeDiscovery.info/

Military Resources
http://www.MilitaryResources.info/

New Economy Analytics, Resources and Alerts
http://www.NewEconomyAnalytics.com/

Outsourcing/Offshoring Information and Resources
http://www.OutsourcingOffshore.us/

Privacy Resources
http://www.PrivacyResources.info/

Reference Resources
http://www.ReferenceResources.info/

Research Resources
http://www.ResearchResources.info/

RestStress™
http://www.RestStress.com/

Script Resources
http://www.WcriptResources.info/

ShoppingBots
http://www.ShoppingBots.info/

Social Informatics
http://www.SocialInformatics.info/

Statistics Resources
http://www.StatisticsResources.info/

Student Research
http://www.StudentResearch.info/

Theology Resources
http://www.TheologyResources.info/

Tutorial Resources
http://www.TutorialResources.info/

World Wide Web Reference
http://www.WWWReference.info/

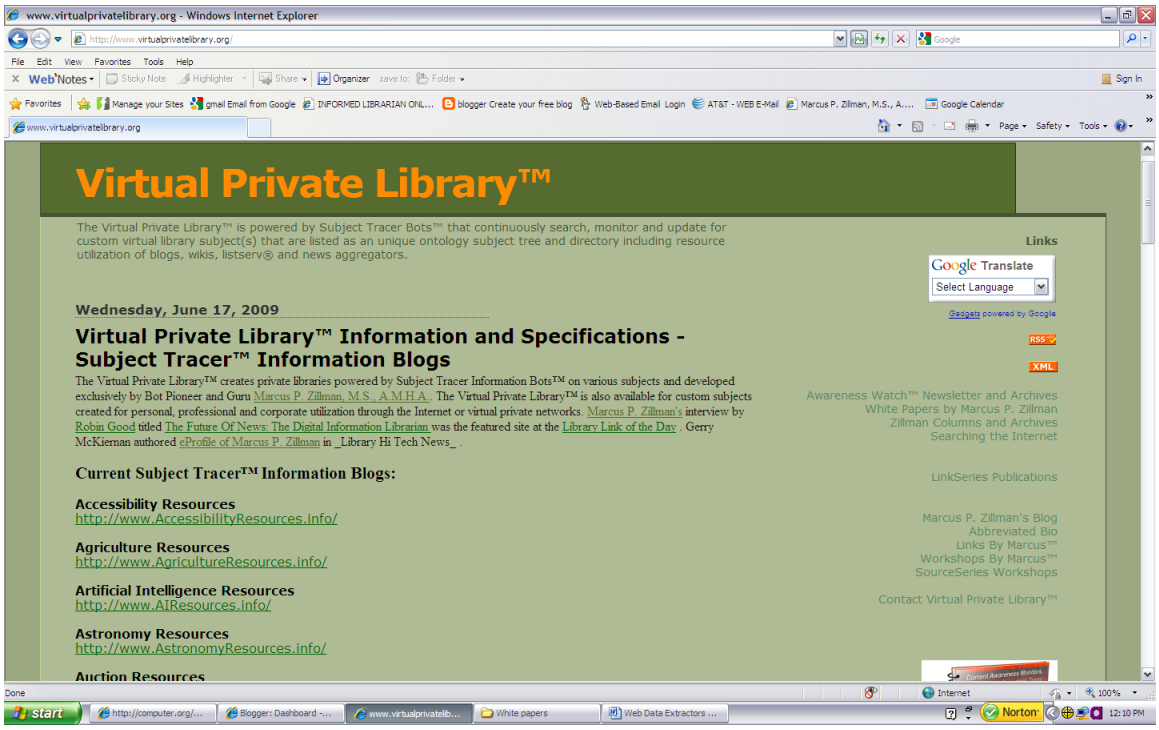

Figure 1: Virtual Private Library™

Author Information: Marcus P. Zillman, M.S., A.M.H.A. Executive Director of the Virtual Private Library is an international Internet expert, author, keynote speaker and corporate consultant in the area of information retrieval, knowledge discovery, knowledge harvesting, artificial intelligence and bots/intelligent agents. He has created numerous world wide web sites including 50 Subject Tracer™ Information Portals and Blogs; written a number of internet miniguides, white papers, manuals and books; hosted over 160 weekly Internet television shows, writes a weekly and monthly column on Current Awareness on the Internet; writes a monthly newsletter Awareness Watch and delivers keynote presentations throughout the international marketplace. He also actively delivers one and two day workshops for key industry sectors displaying how the Internet can be used as a tool to maintain current awareness and professional competencies. Additional websites by Marcus P. Zillman, M.S., A.M.H.A.:

Marcus P. Zillman's Blog
http://www.zillman.us/

Marcus P. Zillman Abbreviated Bio
http://www.zillman.info/

White Papers by Marcus P. Zillman
http://www.WhitePapers.us/

Internet MiniGuides™
http://www.InternetMiniguide.com/

Awareness Watch™ Newsletter
http://www.AwarenessWatch.com/

Marcus P. Zillman's Columns
http://www.ZillmanColumns.com

LinkSeries Publications
http://www.LinkSeries.com/

Internet Sources™ Manual
http://www.InternetSources.info/

Links By Marcus™
http://www.LinksByMarcus.com/

Workshops By Marcus™
http://www.WorkshopsByMarcus.com/

SourceSeries Internet Research Workshops
http://www.SourceSeries.com/

Watch Marcus™
http://www.WatchMarcus.com/

listen to marcus™
http://www.ListenToMarcus.com

Research White Papers, Articles, Lectures and Speeches by Marcus P. Zillman, M.S., A.M.H.A.:

Academic and Scholar Search Engines and Sources
http://www.ScholarSearchEngines.com/

Bots, Blogs and News Aggregators
http://www.BotsBlogs.com/

Business Intelligence Online Resources
http://www.BIOnlineResources.info/

Current Awareness Discovery Tools on the Internet
http://zillman.blogspot.com/2004/09/current-awareness-discovery-tools-on.html

Deep Web Research 2009 Article - LLRX
http://zillman.blogspot.com/2009/01/llrx-december-2008-issue-deep-web.html

eReference Library Link Toolkit
http://www.eReferenceLibrary.com/

Finding Experts By Using the Internet
http://www.FindingExperts.info/

Finding People Resources and Sites
http://www.FindingPeople.info/

Healthcare Bots and Subject Directories
http://www.HealthcareBots.info/

Knowledge Discovery Resources 2009
http://www.KDResources.info/

Online Research Browsers
http://zillman.blogspot.com/2004/10/online-research-browsers-internet.html

Online Research Tools
http://www.OnlineResearchTools.info/

Online Social Networking
http://zillman.blogspot.com/2004/09/online-social-networking-internet.html

Searching the Internet
http://www.SearchingTheInternet.info/

Using the Internet As a Dynamic Resource Tool for Knowledge Discovery
http://zillman.blogspot.com/2004/09/using-internet-as-dynamic-resource.html

Web Data Extractors
http://www.WedDataExtractors.com/

White Papers By Marcus P. Zillman, M.S., A.M.H.A.
http://www.WhitePapers.us/

Internet Tutor by Marcus P. Zillman, M.S., A.M.H.A.
http://www.InternetTutor.info/
Visit this site to learn about the availability of Marcus P. Zillman to tutor you or your associate one on one in the privacy of your residence or office on the latest happenings of the Internet including Internet basics to advanced Internet searching using bots and creating your own personal blog

Internet Speaking by Marcus P. Zillman, M.S., A.M.H.A.
http://www.InternetSpeaker.net
Visit this site to learn about Marcus P. Zillman's speaking engagements for your organization meetings and events. View and listen to his previous presentations as well as his weekly television shows

Internet Consulting by Marcus P. Zillman, M.S., A.M.H.A.
http://InternetConsultant.BlogSpot.com/
Visit this site to obtain information about obtaining the consultation services of Marcus P. Zillman for your company including eCommerce audits, utilization of bots, blogs and news aggregators or the creation of your own personal virtual private library powered by Subject Tracer™ Information bots!

Internet Sources™ Manual
http://www.InternetSources.info
Marcus P. Zillman's latest 378 page manual Internet Sources™ is now available for purchase online and for immediate download. This book makes a great reference resource for the "newbie" to the Internet as well as the seasoned veteran "Internaut".

Current Awareness Monitors, Alerts and Information Traps for 2008
http://www.ecurrentAwareness.com/
Marcus P. Zillman's latest report Current Awareness Monitors, Alerts and Information Traps for 2008 is now available for purchase online and for immediate download. This report is a comprehensive listing of the latest resources, sources and sites for current awareness on the Internet. This is a must read for anyone who must stay current in their profession and/or business activity as the list of URLs will keep you at the leading edge of your career.

Market Intelligence Resources 2008
http://www.MarketIntelligenceResources.com/
Marcus P. Zillman's just released professional Internet MiniGuide is titled Market Intelligence Resources 2008 and is now available for purchase online and immediate download. This 130 page digital miniguide represents a comprehensive listing of the latest resources, sources and sites to discover the latest Market Intelligence sources available on the Internet with many of them freely available! Designed specifically for today's entrepreneur, professional and/or investor.

Entrepreneurial Links 101
http://www.EntrepreneurialLinks.com/
Marcus P. Zillman's newly released 130 page eReference digital book for the up and coming entrepreneur. Entrepreneurial Links 101 gives an alphabetical listing of the very best Internet and World Wide Web sites covering Entrepreneur Resources, Business Intelligence Resources and an extremely comprehensive list of Online Research Tools. This is considered by many to be the entrepreneur's bible for finding relevant and competent online resources!

Internet Privacy and Security Resources
http://www.InternetPrivacySecurity.net/
Marcus P. Zillman's latest eReference digital publication is a selected comprehensive alphabetical listing of the latest resources and sites covering all aspects of privacy and security currently available over the Internet. From the board room to the family room, these resources and sites give you the information you need to maintain your privacy and security as you use the Internet in your business and personal life.

Research Resources Online Guide
http://www.ResearchResourcesOnline.net/
Marcus P. Zillman's latest LinkSeries Publication is a 170 page digital guide of a selected comprehensive alphabetical listing of the latest and greatest resources and sites covering all areas of research that is currently available over the Internet. The guide covers online research resources and tools for the Newbie to research as well as the Seasoned researcher. Contents include: a) Research Resources, b) Research Tools, c) Student Research Resources Toolkit, d) Knowledge Discovery/Management and Data Mining Resources, e) Knowledge Discovery/Retrieval and the World Wide Web Resources, and f) Subject Tracer™ Information Blogs.

The Survivor's Manual for The New Economy.
http://www.NewEconomyManual.com/
Marcus P. Zillman's latest LinkSeries Publication is a 239 page digital read that gives excellent resources and annotated sources for the new economy analytics, alerts,

ecommerce, financial sources, invisible and deep web resources, social and business networking sources along with new economy competitive and business intelligence resources and an extremely comprehensive listing of new economy online tools.